AUTHORS' RESPONSES TO REVIEWERS COMMENTS FOR "META-LEARNING HOW TO FORECAST TIME SERIES"

Manuscript number FOR-21-0572

We would like to thank the Editor, Associate Editor and the two anonymous reviewers for reviewing our paper. We also thank you for the insightful comments and suggestions. In this revision, we have addressed all the comments raised by the reviewers, providing a point-to-point response to each comment made by the review team.

Responses to Referee(s)' Comments

First, we would like to thank the reviewer for the time of reviewing our paper. We have tried our best to address all of your comments and suggestions. Below we provide a point-to-point response to your comments.

Reviewing: 1

The subject of the work is important, the work is mature, well written, and the experimental part is extensive. To improve further the paper quality, the Authors can take into account the comments below.

Detailed comments:

(1) Page 7/line 5: what do you mean by the test set? Usually the test set (separate from the training and validation sets) cannot be used for training, optimization and selection of the model. Clarify, please.

We thank the reviewer for raising this concern.Unfortunately, we were unable to find the place that corresponds to this comment. However, we have clarified this point in the revised manuscript in Section 2.3 and highlighted it below.

The training set is used to estimate the parameters of a forecasting model. Based on this fitted model, we generate forecasts over the test set and compare them against the test set values. Since the test data are not used for model fitting, this practice provides a reliable indicator to evaluate how well the model makes accurate forecasts on new data.

(2) Fig 3. It looks like red "feature calculation" block does the same work as blue "input-features" block. Thus unify the names of these blocks so as not to cause confusion.

We thank the reviewer for the suggestion. We changed the wording in Fig 3 - red part "feature calculation" to "input-features" to unify the names and avoid confusion.

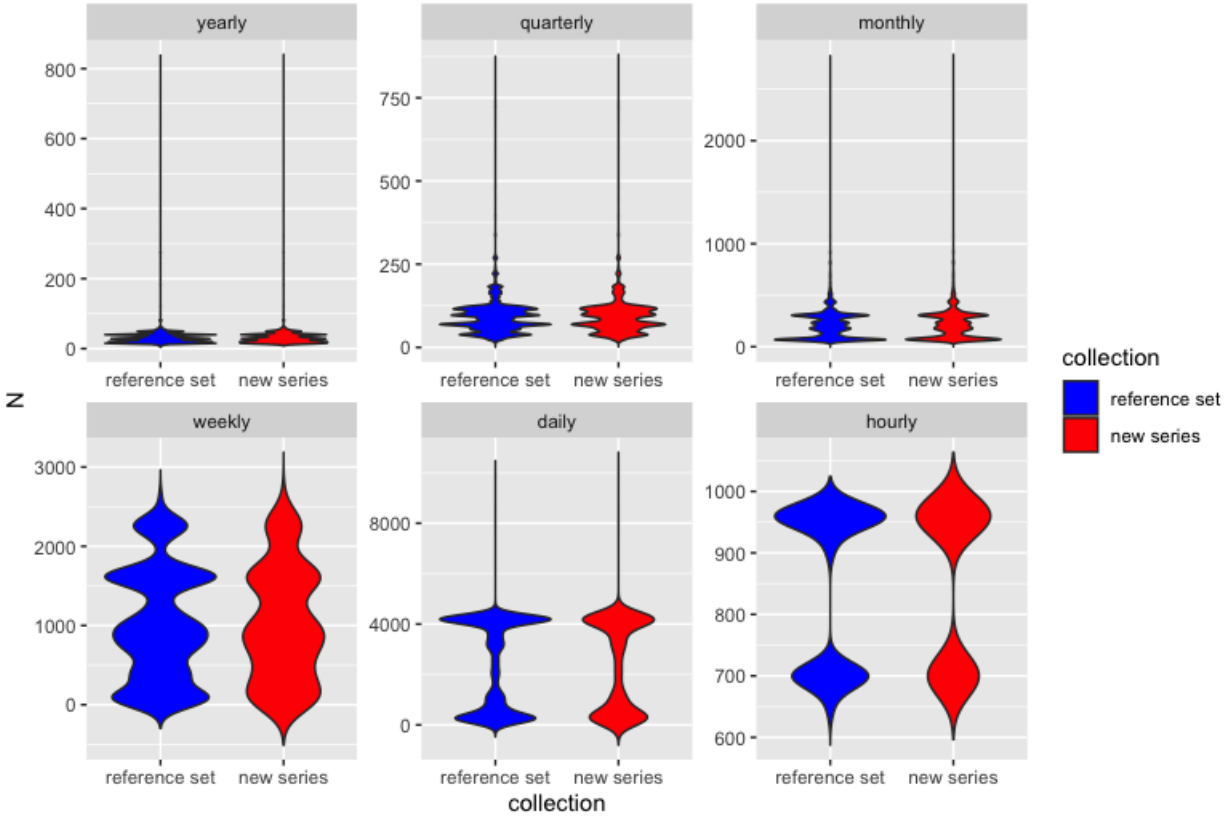
(3) 9/54: Explain how the time series are split into training and test part. How sensitive is the proposed approach to this split?

Thank you for the comment. We include length as one of our features, but the remaining features are asymptotically independent of the length of the time series (i.e., they are ergodic), and they are independent of scale. As our main focus is forecasting, we select features which have discriminatory power in selecting a good model for forecasting. This point is mentioned in the last paragraph of section 2.1. In addition to that we added following to the requested section

Usually, the first 80% of a time series is used as the training set, and the remaining last 20% is used as the test set for evaluation. In our research, since we used the M-competion data, we used the training and test set specified by the M-competion organizers. In the M-competitions training set length for series varies, however the test length for the time series are fixed for each frequency level. The number of observations in the test periods are 6 for yearly, 8 for quarterly, 18 for monthly, 13 for weekly, 14 for daily and 48 for hourly series.

Use of this standard test set specified by the M-competitions allows to compare our results with the results of other studies. Figure *** provides the distributions of length of the training period of the M-competitions and simulated data used in the study.

To show the distribution of length in the reference set and new series collection we added the following figure



Furthermore, we added following to the conclusions section

Our method is sensitive to the training duration of the time series because we employed length as an input feature. We found that whereas longer time series prefer to choose more parameterized models, shorter time series are more likely to choose simple models like random walk, white noise, etc. However, PDP curves of length revealed that there is a threshold beyond which the probability of choosing various models in response to training set length starts to plateau.

(4) 9/55: Fitting the model is related to its training and optimization (selection of hyperparameters). Moreover, the model performance depends on the feature selection/engineering. I suppose that these all tasks are performed by the "fit models" block in Fig. 3. You omit the topic of model optimization including feature selection/engineering, which can be very complex and time consuming process due to huge search space for same models. Describe this topic to make the reader aware of its importance.

(5) Algorithm 1: One of the key hyperparameters of random forests, in addition to the number of trees and the number of features to be selected at each node, is the minimum number of leaf node observations or its equivalent. You don't mention this hyperparameter at all. Why?

(6) 11/5: "One approach is to fit models to the time series in the reference set, and then use those models as data generating processes to obtain similar time series." Clarify this issue.

(7) Time series augmentation looks like important component of the proposed framework and should be described in detail.

(8) 13/46: "incorporating class priors into the RF classifier". Could you explain this, please.

(9) Many forecasting models, especially machine learning ones such as NNs, have a stochastic nature, which translates into different results for the same input data in different training sessions. This may have a negative impact on the choice of the best model. How does your proposed approach deal with such problem? Explain, please.